SEQUENCE MATCHING, SIMPLE SEARCHING

PGA Course in Bioinformatics
Tools for Comparative Analysis

February 24, 2003

Outline

- ❧Sequence alignment algorithms
  - Rigorous Optimality:Needleman-Wunsch and Smith-Waterman
  - Rapid, heuristic algorithms
    - BLAST
    - FASTA
    - and their relatives
- ❧Databases and Search Tools

MAJOR SITES WE WILL USE

- http://www.ncbi.nlm.nih.gov/
- http://workbench.sdsc.edu

What are you Comparing

- **Homologue**

  Sequences that share a common ancestor; may have similar function
- **Paralogue**

  Similar sequence within species, may have similar function
- **Orthologue**

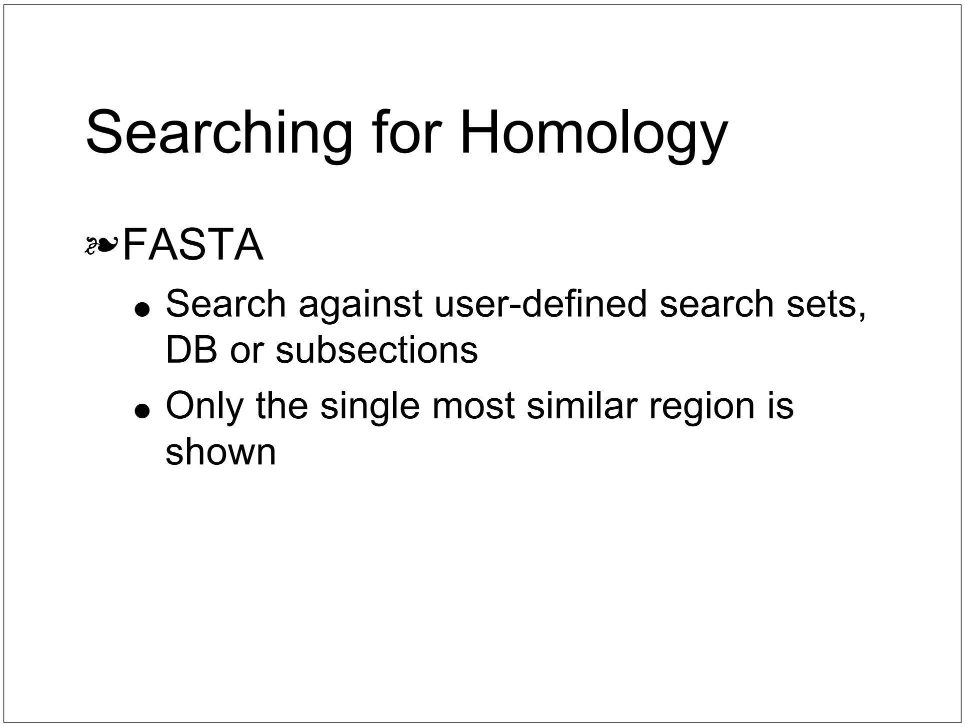
  Same sequence separated by a speciation event, probably same function

ANALOG

Non-homologue proteins that have similar folding architecture, or similar functional sites, which are believed to have arisen through convergent evolution

Searching for homology

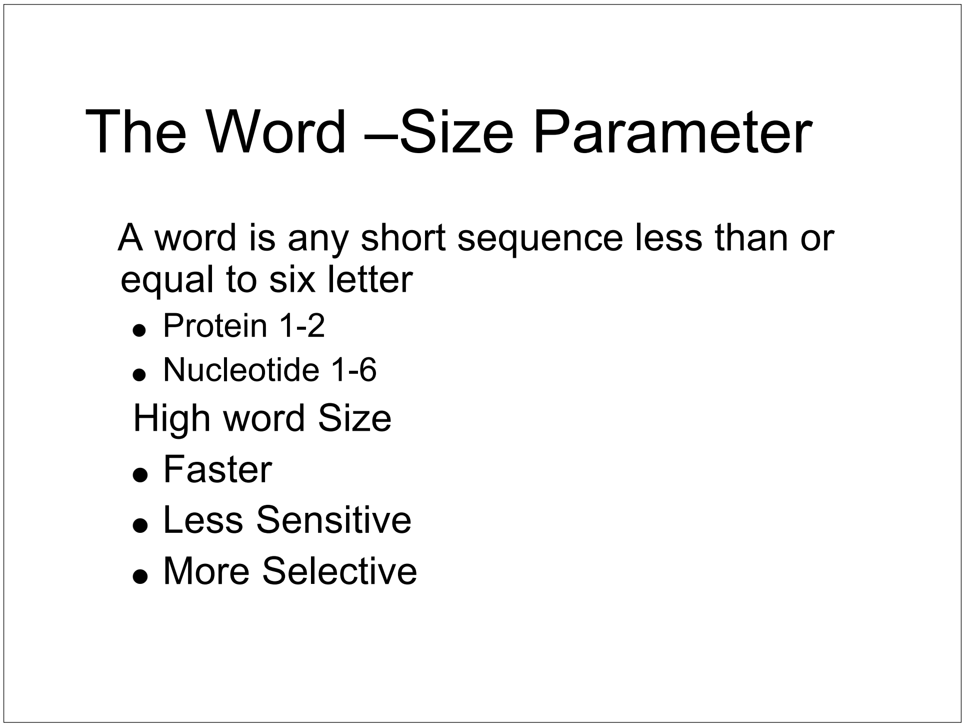
- BLAST
  - Remote search at NCBI or locally
  - Non-redundant set of databases, one DB at a time
  - Fast
  - Shows several similar regions
  - Less sensitive for (shorter) nucleotide sequences



Searching for Homology

❧FASTA

- Search against user-defined search sets, DB or subsections
- Only the single most similar region is shown



The Word –Size Parameter

A word is any short sequence less than or equal to six letter

- Protein 1-2
- Nucleotide 1-6

High word Size

- Faster
- Less Sensitive
- More Selective

Evolution and Alignment

Evolutionary concepts enable the determination of similarity and homology

- Similarity is an observable quantity, such as %identity
- Homology is a conclusion drawn from the data that two genes share a common evolutionary history.

Evolution and Alignments (2)

- Genes are either homologous or not homologous.
- There is no degree of homology
- You can’t tell what the ancestral sequence is simply because you have two or more homologues.

So, what IS an Alignment?

Evolution and Alignments (3)

- Alignments reflect the PROBABLE evolutionary history of two sequences
- Residues that align and are not identical represent substitutions
- Sequences without correspondence is aligned sequences are interpreted as indels and in an alignment are gaps.

Evolution and Alignment

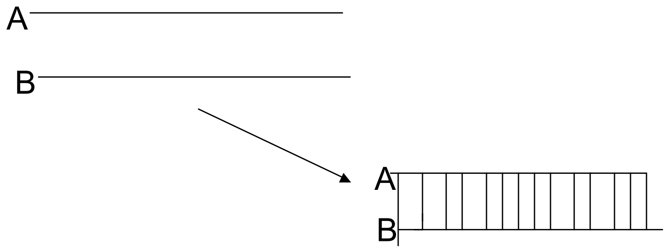
- Certain regions are more conserved than others, based on structure/function
- Certain regions may be conserved simply by history, not function
- This is true especially for closely related species.

Structure and Alignment

- If two proteins have more than 20-30% ID aligned, then the 3-D structures tend to be similar
- Overall folds are the same, details differ
- Form often follows function (Beware the BUT).
- So, sequence alignment is sometimes a 3-D alignment.

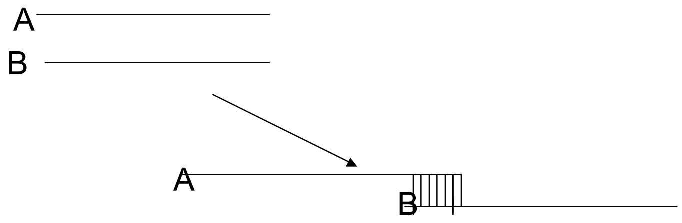
Global Alignment

Optimal alignment over the entire length



Local Alignment

Finds the highest coring alignment regardless of position and length



Needleman Wunsch Algorithm

- Global alignment:: every residue of the two sequences has to participate
- Guaranteed to calculate an Optimal similarity score
- Begin at the beginning of each sequence and go to the end.
- Cannot detect domains

Smith-Waterman Algorithm

- Optimal Local Alignment
- Guaranteed to find all significant matches to a given query
- Takes the query sequence versus every sequence in the database
- Can be used with arbitrary scoring systems
- **COMPUTATIONALLY EXPENSIVE!!!**

Scoring Matrices

- Relatively simple for DNA-gap penalties or mismatches-can be made to look at Pu/Py

- Protein matches look also at similarity (leu/ileu)

Protein Scoring Matrices

- ❧Chemical similarity: 210 pairs of aa
- ❧Nearness in Genetic Code
- ❧Chemical similarity, e.g., hydrophobicity
- ❧Observed Substitution Schemes

AA Substitution Matrices

Rationale:

Certain amino acid substitutions commonly occur in related proteins (sometimes from different species). These provide the basis for amino acid substitution matrices, essentially a symbol comparison table.

More on Matrices

- A substitution matrix specifies a set of scores s_{ij} for replacing amino acid I by amino acid j.

- PAM: Percent Accepted Mutations
- BLOSUM Blocks Amino Acid Substitution Matrices

Amino Acid Symbols

- A Ala alanine
- B   Asx Aspartic or asparaine
- C   Cys Cysteine
- D   Aspartic acid
- E   Glu Glutamic acid
- F   Phe Phenylalanine
- G Gly Glycine
- H   His Histidine
- I   Ile Isoleucine
- K   Lys Lysine
- L   Leu Leucine
- M Met Methionine
- N Asn Asparagine
- P Pro Proline
- Q Gln Glutamine

R Arg Argine
S Ser Serine
T Thr Threonine
U Sec Selenocysteine
V Val Valine
W Trp Tryptophan
X   Xaa Unknown or other aa
Y   Tyr Tyrsoine
Z   Glx Glutamic or glutamine

Observed AA Substitution Matrices

- PAM
- BLOSUM

PAM

- Log Odds scores are used
- The score of each pair s(a,b) is defined as the log of the likelihood ratio of the transition probability M_{ab} (Mutation) versus the probability of a random occurrence of the amino acid b in the second sequence.

$$s(a,b)=\log M_{ab}/P_b$$

PAM: Point Accepted Mutation

- DAYHOFF et al.
- Observed residue replacement in related proteins
- GLOBAL alignment, closely related
- A model of molecular evolution

1 PAM = average change in 1% of all amino acid possibilities(1% divergence)

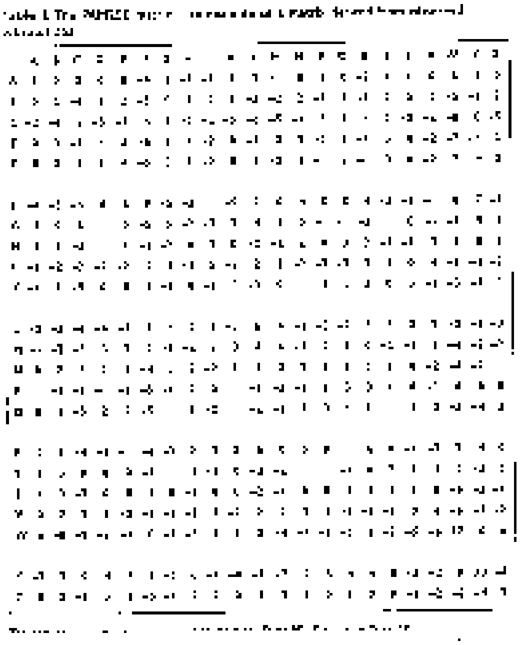
- Other PAM matrices extrapolated from PAM1.

PAM continued

- TIME is NOT correlated with PAM
- Number of the matrix refers to evolutionary distance

Means different families of proteins evolve at different rates

PAM250



[illegible]

BLOSUM

- Block Substitution Matrix
- Henikoff and Henikoff, PNAS, 1992
- Number following indicates per cent identity within set, BLOSUM62=62% id
- Finds short, highly similar sequences (no gaps)

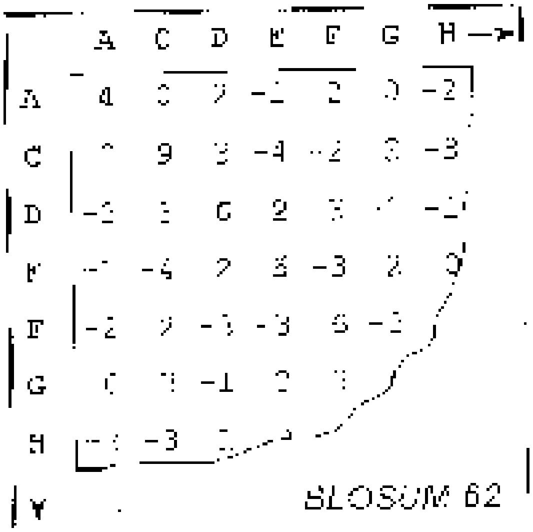
BLOSUM

- Matrices are directly calculated, based on observed alignments
- Greater numbers are lesser distances
- Usually best for local similarity searches
- BLOSUM62= DEFAULT FOR BLAST. If a distant relative, think about another matrix.

BLOSUM SCORING RULES

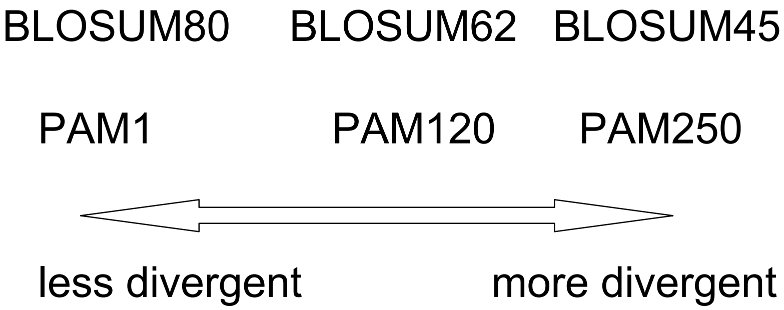
- Zero score means the frequencies of the pair in the database is that expected by chance
- A positive score means more frequent than chance
- Negative score means the pair is found less frequently than chance.

Blosum62



	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	
G	0	-3	-1	-2	-3		
H	-2	-3	-1				
Y							

BLOSUM 62



BLOSUM80 BLOSUM62 BLOSUM45

PAM1 PAM120 PAM250

less divergent more divergent

BLAST-Basic Local Alignment Sequence Tool

- Objective: find all local regions of similarity distinguishable from random
- Only local alignments permitted,
- Gaps permitted in version 2
- Statistically sound (Karlin and Altschul), but no guantee of optimality

BLAST: Three Step Algorithm

- Compile a list of high scoring words of length w (w=4 for proteins, 12 for nucleic acids)
- Scan for word hits of score greater than threshold, T
- Extend word hit in both directions to find High Scoring Pairs with scores greater than S

Other BLAST Programs

- BLASTN: nucleic acid query to NA database
- BLASTP: Protein query to Protein database
- BLASTX: Translated nucleic acid query to Protein database
- TBLASTN: Protein query against (translated) nucleic acid database
- TBLASTX: Translated nucleic acid against translated nucleic acid database

OTHER BLAST VARIATIONS

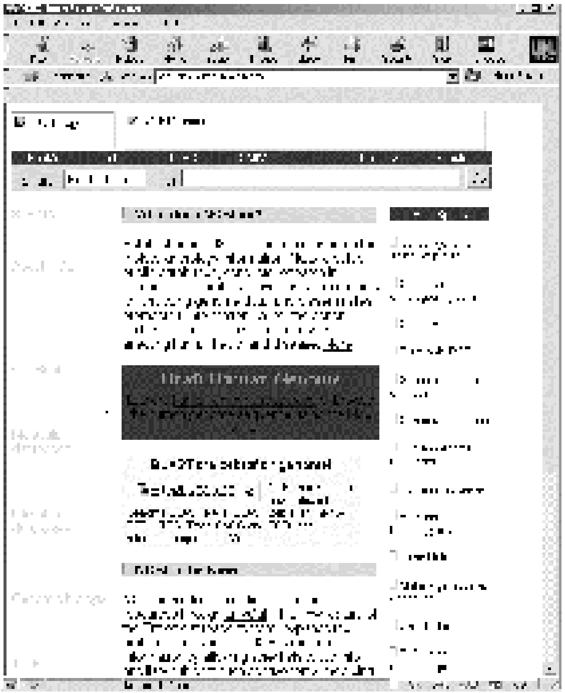
- Gapped BLAST (BLAST 2.0) -extend words from no-gap to gap, generate gapped alignments
- PSI-BLAST- Position Specific Iterated BLAST-use gapped BLAST, generate a Profile from multiple iterations used instead of the input and Distance Matrix

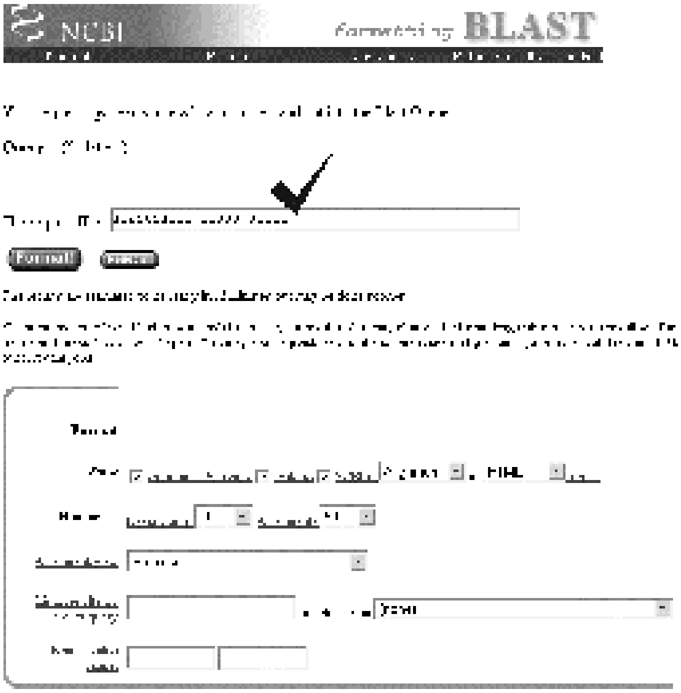
Limitations to BLAST

- Needs islands of strong homology
- Limits on the combination of scoring and penalty values
- The variants (blastx, tblastn, tblastx) use 6-frame translation-miss sequences with frameshifts)
- Finds and reports ONLY local alignments

A WALK THROUGH BLAST

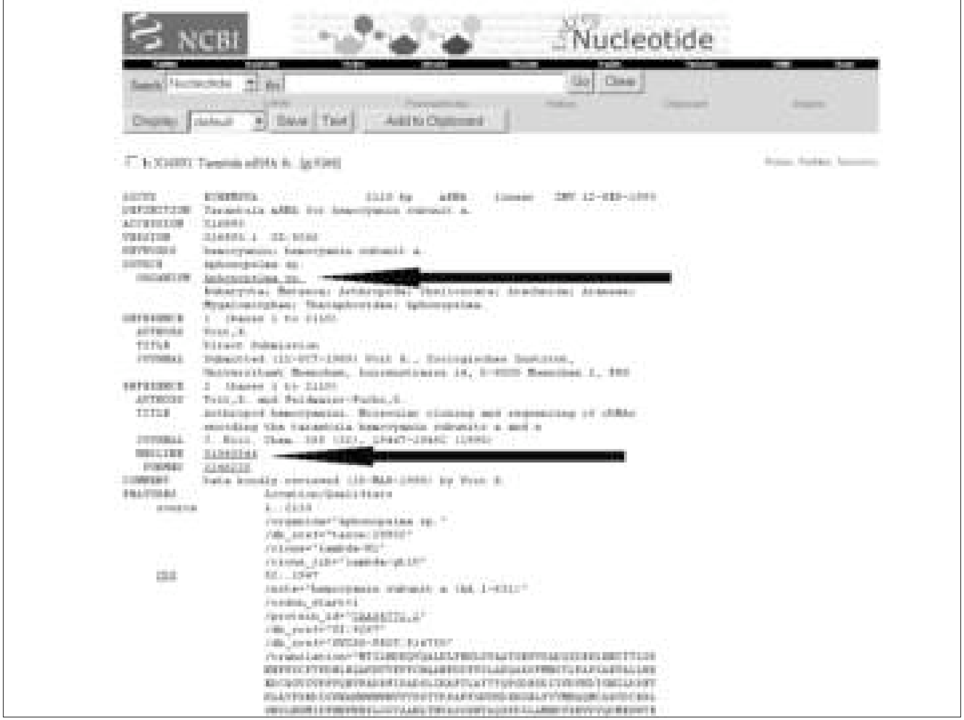
NCBI home





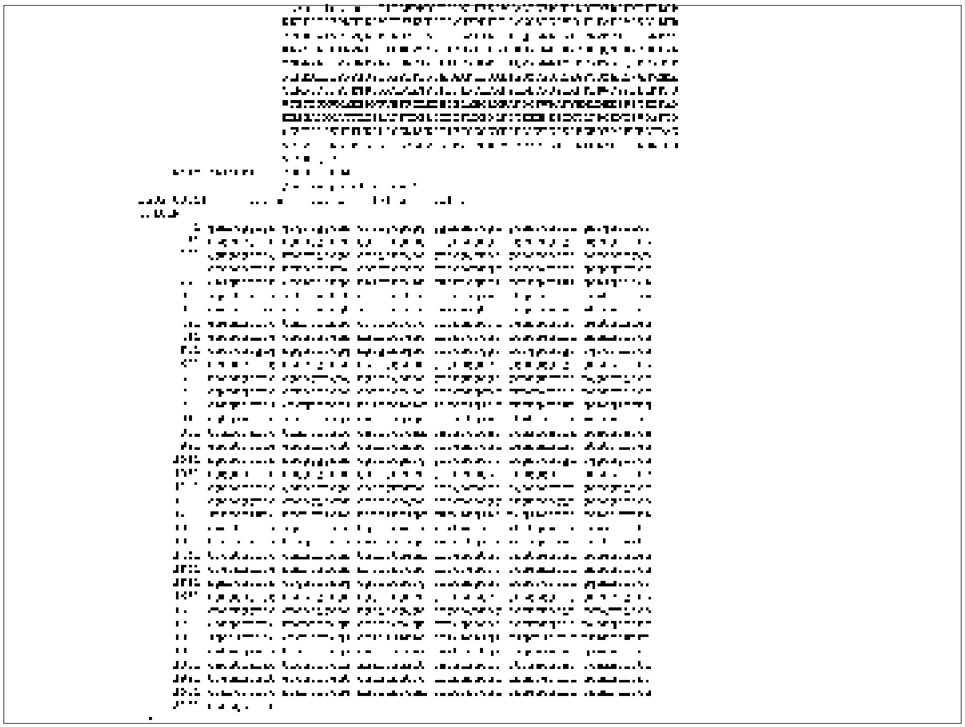
NCBI
results of BLAST

NCBI
results of BLAST



□ 1: [illegible] Tarantula mRNA fo... [gi:[illegible]]

```
LOCUS       [illegible]              2110 bp    mRNA    linear   INV [illegible]
DEFINITION  Tarantula mRNA for hemocyanin subunit a.
ACCESSION   [illegible]
VERSION     [illegible]  GI:[illegible]
KEYWORDS    hemocyanin; hemocyanin subunit a.
SOURCE      Aphonopelma sp.
  ORGANISM  Aphonopelma sp.
            Eukaryota; Metazoa; Arthropoda; Chelicerata; Arachnida; Araneae;
            Mygalomorphae; Theraphosidae; Aphonopelma.
REFERENCE   1  (bases 1 to 2110)
  AUTHORS   Voit,R.
  TITLE     Direct Submission
  JOURNAL   Submitted ([illegible]) Voit R., Zoologisches Institut,
            Universitaet Muenchen, Luisenstrasse 14, D-8000 Muenchen 2, FRG
REFERENCE   2  (bases 1 to 2110)
  AUTHORS   Voit,R. and Feldmaier-Fuchs,G.
  TITLE     Arthropod hemocyanins. Molecular cloning and sequencing of cDNAs
            encoding the tarantula hemocyanin subunits a and e
  JOURNAL   J. Biol. Chem. [illegible]
  MEDLINE   [illegible]
   PUBMED   [illegible]
COMMENT     Data kindly reviewed ([illegible]) by Voit R.
FEATURES             Location/Qualifiers
     source          1..2110
                     /organism="Aphonopelma sp."
                     /db_xref="taxon:[illegible]"
                     /clone="[illegible]"
                     /clone_lib="[illegible]"
     CDS             [illegible]
                     /note="hemocyanin subunit a [illegible]"
                     /codon_start=1
                     /protein_id="[illegible]"
                     /db_xref="GI:[illegible]"
                     /db_xref="[illegible]"
                     /translation="[illegible]
```



BLAST RULES OF THUMB

- For short amino acid sequences (20-40), 50% identity happens by chance
- If A and B are homologous, and B and C are homologous, then A and C are, even if you can’t see it.
- You can get similarity in the absence of homology for low complexity, transmembrane and coiled-coil regions. These have to be eliminated by you, but you MAY want them.

BLAST Significance

- If you change scoring systems, you can still compare search results if you normalize the score.

  S’=(lambdaS-lnK)/ln2. Lambda and K are associated with the scoring system.

  S’, with a given E, is significant if it is greater than log N/E, N the size of the search space.

FASTA: WHY USE IT?

- Allow alignments to shift frames

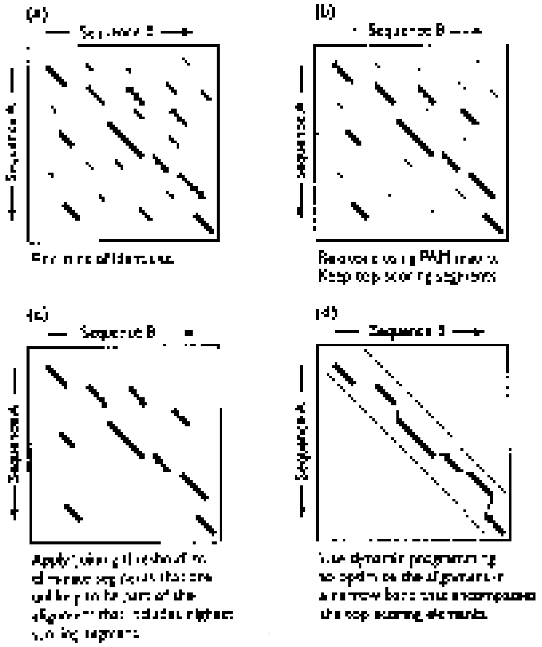
FASTA: FAST Alignment

- http://alpha10.bioch.virginia.edu/fasta/
- http://www2.ebi.ac.uk/fasta3
- http://workbench.sdsc.edu

- Rapid Global alignment

- Not a strong mathematical basis

FASTA Algorithm



(a) Find runs of identities

(b) Rescore using PAM matrix. Keep top scoring segments

(c) Apply joining threshold to eliminate segments that are unlikely to be part of the alignment that includes highest scoring segments

(d) Use dynamic programming to optimize the alignment in a narrow band that encompasses the top scoring elements

LALIGN

- Essentially a FASTA derivative for local alignments
- Compares two proteins to identify regions of similarity
- Will report several sequence alignments within a given sequence
- Works for internal repeats that are missed by FASTA because of gaps.

SITEs for LALIGN

- http://fasta.bioch.virginia.edu/fasta/lalign.htm
- http://xylian.igh.cnrs.fr/bin/lalign-guess.cgi
- http://biowb.sdsc.edu (registration necessary but painless)
- PALIGN http://fasta.bioch.virginia.edu/fasta/palign.htm (plots a graph of the areas of alignment)

ENTREZ: Linked Databases

http://www.ncbi.nlm.nih.gov/Entrez/

- Concept of Neighbor-usually BLAST relationship
- Precomputed=Fast
- Related sequence, structure neighbors, related articles

- CUBBY

EST DATABASES:Quality issues

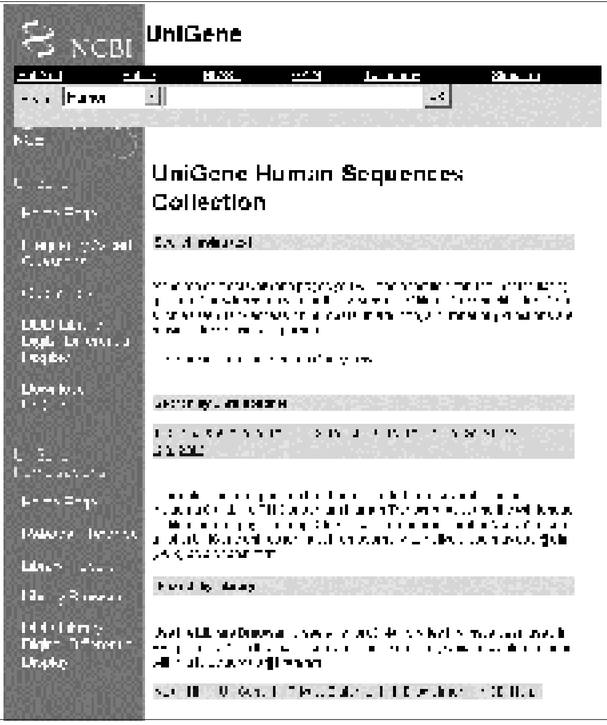
- SEQUENCE QUALITY
  - calculated error less than 1% (Phred-20) is the rule
  - frameshifts and stops common
  - Rules are usually observed by exception
  - There are lots of exceptions in the public data
  - Many 3’ UTRs

EST Databases: Quality #2

- CLONE QUALITY
  - Over-representation
  - Tissue specificity
  - Developmental stage specificity
  - Unprocessed mRNA clones
  - Chimeras
  - Contamination

EST Cluster Databases

- STACK-at SANBI http://sanbi.ac.za
- TIGR-animals, plants, other http://www.tigr.org/tdb/tgi.shtml
- Unigene-NCBI
  - Human, mouse, rat, cow,zebrafish
  - mRNAs
  - predicted mRNAs



UniGene

UniGene Human Sequences Collection

UNIGENE

- A LIST OF LISTS
  - The cluster and known EST, mRNA pieces
  - Additional annotation-gene name, etc.
  - Distributed as a subset of dbest

NOT included in the BLAST searchable DB at NCBI

Caveats on Clusters

- Not stable
- Can go to complete cDNAs as available

LOCUSLINK

(http://www.ncbi.nlm.nih.gov/LocusLink)

- A useful, searchable compendium of loci across human, mouse, rat, Drosophila and zebrafish
- Linked for PubMed, OMIM, RefSeq, Homologene data, Unigene, and Variation Data

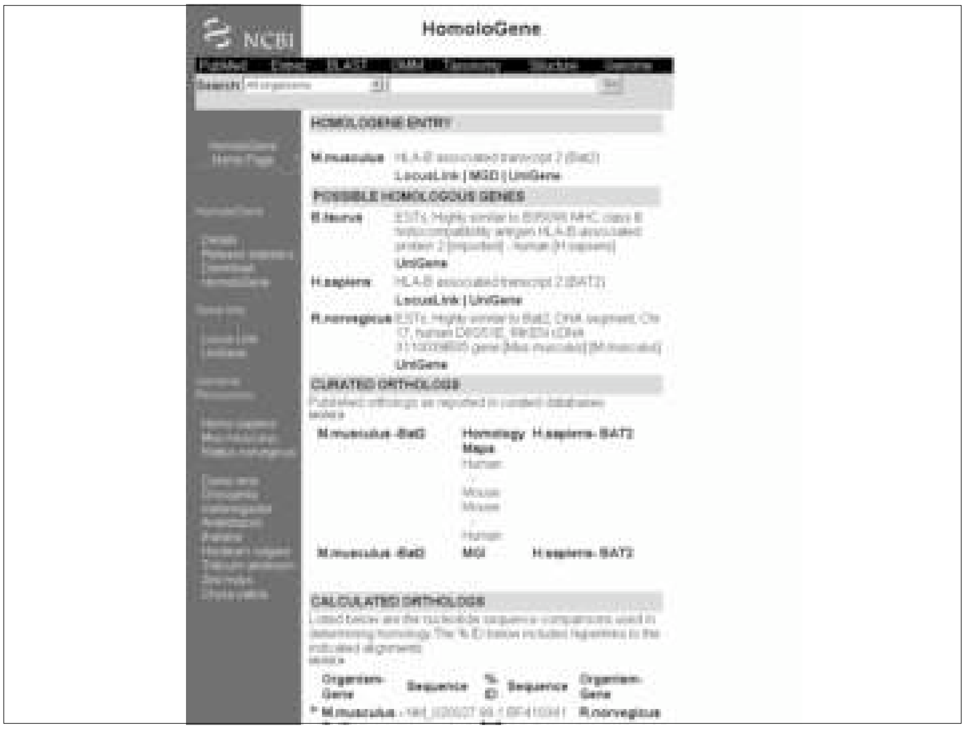
LocusLink provides links into:
Pubmed,
Omim,
Refseq,
HomolGene,
Unigene,
Variation data

HomoloGene

[illegible]

HomoloGene [illegible]

[illegible]



HomoloGene

HOMOLOGENE ENTRY

M.musculus HLA-B associated transcript 2 (Bat2)
LocusLink | MGD | UniGene

POSSIBLE HOMOLOGOUS GENES

B.taurus ESTs, Highly similar to [illegible] MHC class III histocompatibility antigen HLA-B associated protein 2 [imported] - human [H.sapiens]
UniGene

H.sapiens HLA-B associated transcript 2 (BAT2)
LocusLink | UniGene

R.norvegicus ESTs, Highly similar to Bat2, DNA segment, Chr 17, human D6S51E, RIKEN cDNA [illegible] gene [Mus musculus] [M.musculus]
UniGene

CURATED ORTHOLOGS

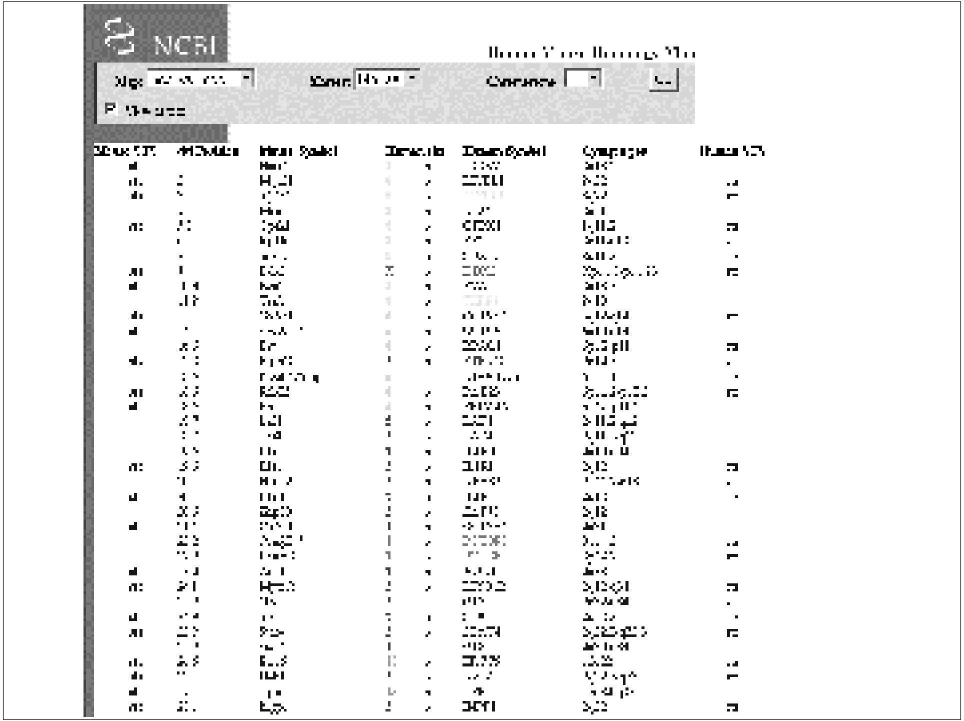
Published orthologs as reported in curated databases

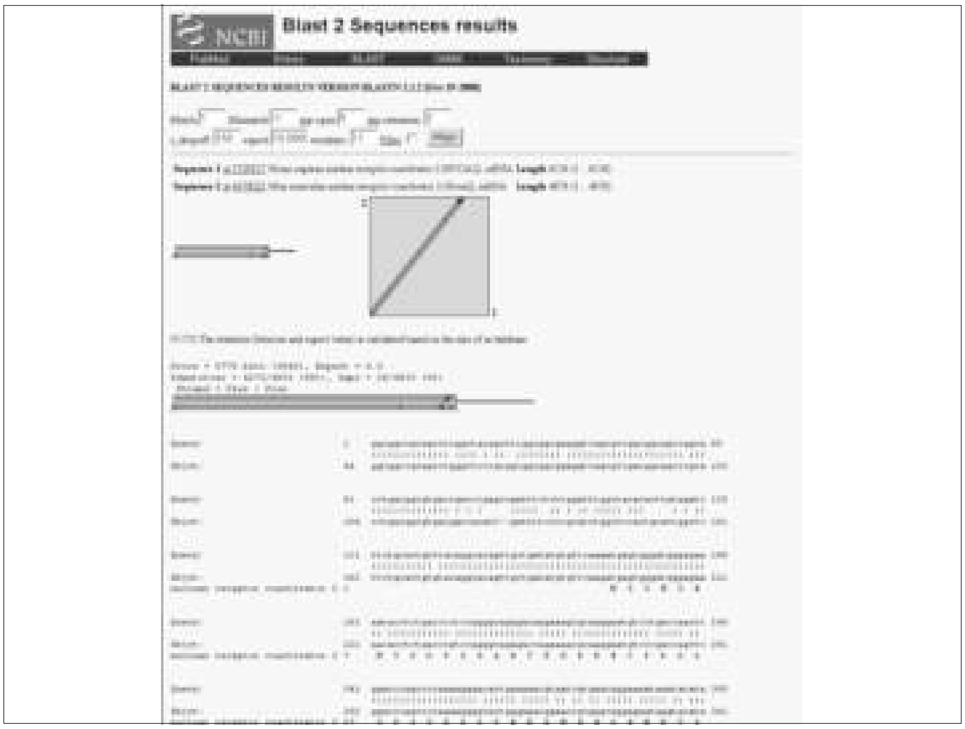
M.musculus -Bat2	Homology Maps Human - Mouse Mouse - Human	H.sapiens- BAT2
M.musculus -Bat2	MGI	H.sapiens- BAT2

CALCULATED ORTHOLOGS

Listed below are the nucleotide sequence comparisons used in determining homology. The % ID below includes hyperlinks to the indicated alignments

	Organism-Gene	Sequence	% ID	Sequence	Organism-Gene
*	M.musculus	NM_[illegible]	99.1	BF[illegible]	R.norvegicus





Blast 2 Sequences results

Resources for Genomic Comparison

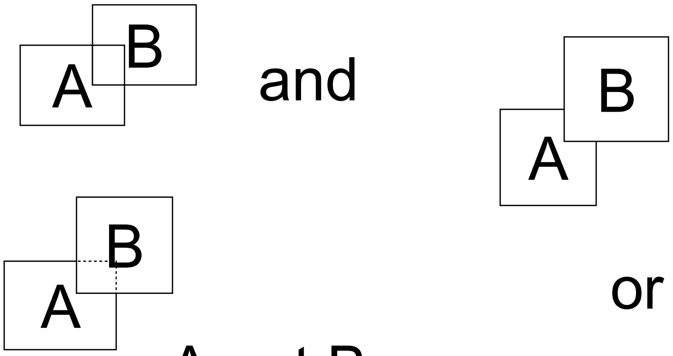
- GLASS-http://plover.lcs.mit.edu
- PipMaker: http://bio.cse.psu.edu
- Rosetta: http:// plover.lcs.mit.edu(genes)
- SGP: htttp://soft.ice.mpg.de/sgp-1
- VISTA: http://www-gsd.lbl.gov/VISTA
- WABA: http://www.cse.ucsc.edu/~kent/xenoAli/index.html

EFFICIENT SEARCHING

- Use Wild Cards: #,$,?,*
- Use Boolean Operators
  - Not
  - And
  - Or
  - Nor

Boolean Operators

- *AND* A and B BOTH
- *OR* A or B EITHER
- *NOT* B not A Have B, do not have A
- *NOR* A nor B A but not B OR B but not A



and

or

A not B

WILD CARDS

- Match one character-NCBI uses #
- Match zero or one character NCBI uses $, others ?
- Match zero or more characters-usually *

RULES OF THUMB

- Use an up-to-date database; repeat often
- Choose a fast algorithm
- Use the most recent version
- Work at the protein level--for a small amount of evolutionary change, DNA sequence contains less information about homology
- Respect your own ***intuition***

MEDICAL SUBJECT HEADINGS

❧CONTROLLED Vocabulary

❧Indexing of articles, books, etc.

❧Current version has over 300,000 terms

❧Can download list and make your own assortment

MeSH Advantages

❧Assigned to the the entire document, not just title and abstract

❧Major topic (*)

❧Subheadings if available

❧MeSH topics are exploded to include all the terms included in the meaning.

Try it; you may like it.

Gene Ontologies GO

A gene ontology is a controlled vocabulary used to describe the biology of a gene product in any organism, designed to allow both attribution and querying at different levels of granularity , facilitating queries across participating databases.

A step toward unifying biological databases but not sufficient.

http://www.geneontlogy.org

Components of GO

A gene product is a physical thing (protein, RNA, can have small molecules associated to make a gene product group.

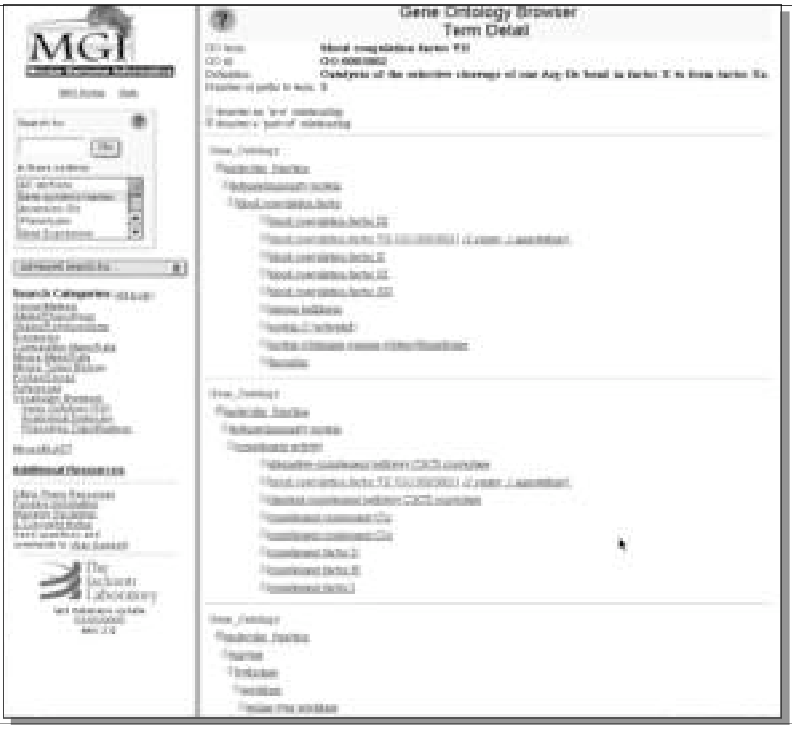
Attributes of Gene Products

- **Molecular Function**-what something does
- **Biological process**-a biological objective, like growth or pyrimidine metabolism
- **Cellular Component-**part of a cell, ER, nucleus etc.

Ontology Representations

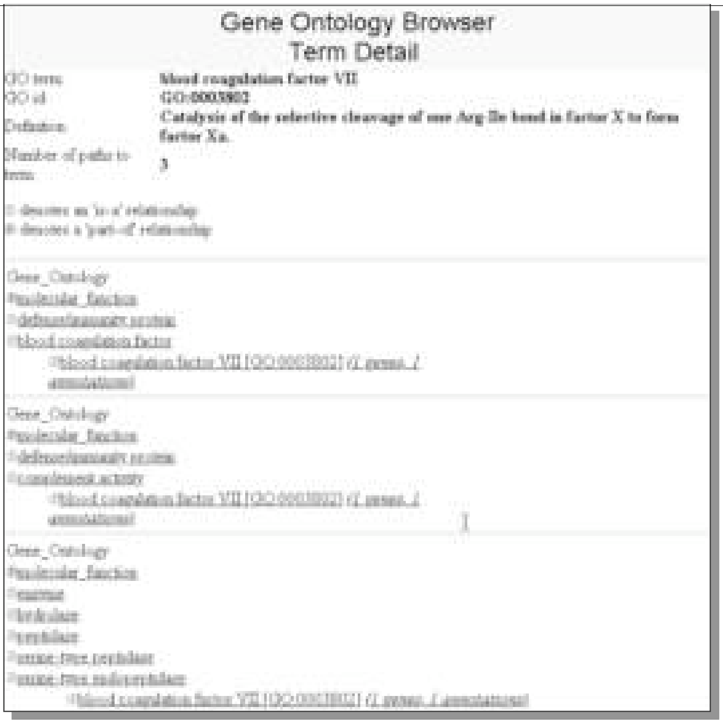
- A network, a directed acyclic graph (DAG), in which terms have multiple parents and multiple relationships to parents.
- Relationships connecting terms include is-a, part-of,

Yeast, Fly, Mouse, Arabidopsis, Worm,



Gene Ontology Browser

Term Detail



Gene Ontology Browser
Term Detail

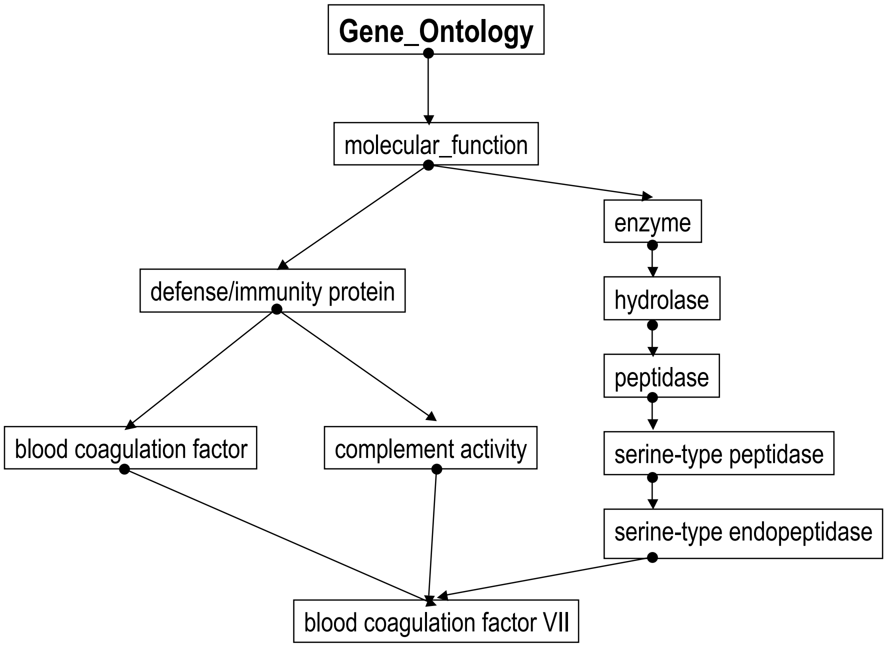
GO term:	**blood coagulation factor VII**
GO id:	**GO:0003802**
Definition:	**Catalysis of the selective cleavage of one Arg-Ile bond in factor X to form factor Xa.**
Number of paths to term:	**3**

denotes an 'is-a' relationship
denotes a 'part-of' relationship

Gene_Ontology
molecular_function
defense/immunity protein
blood coagulation factor
blood coagulation factor VII [GO:0003802] (1 genes, 1 annotations)

Gene_Ontology
molecular_function
defense/immunity protein
complement activity
blood coagulation factor VII [GO:0003802] (1 genes, 1 annotations)

Gene_Ontology
molecular_function
enzyme
hydrolase
peptidase
serine-type peptidase
serine-type endopeptidase
blood coagulation factor VII [GO:0003802] (1 genes, 1 annotations)



Gene_Ontology

molecular_function

enzyme

defense/immunity protein

hydrolase

peptidase

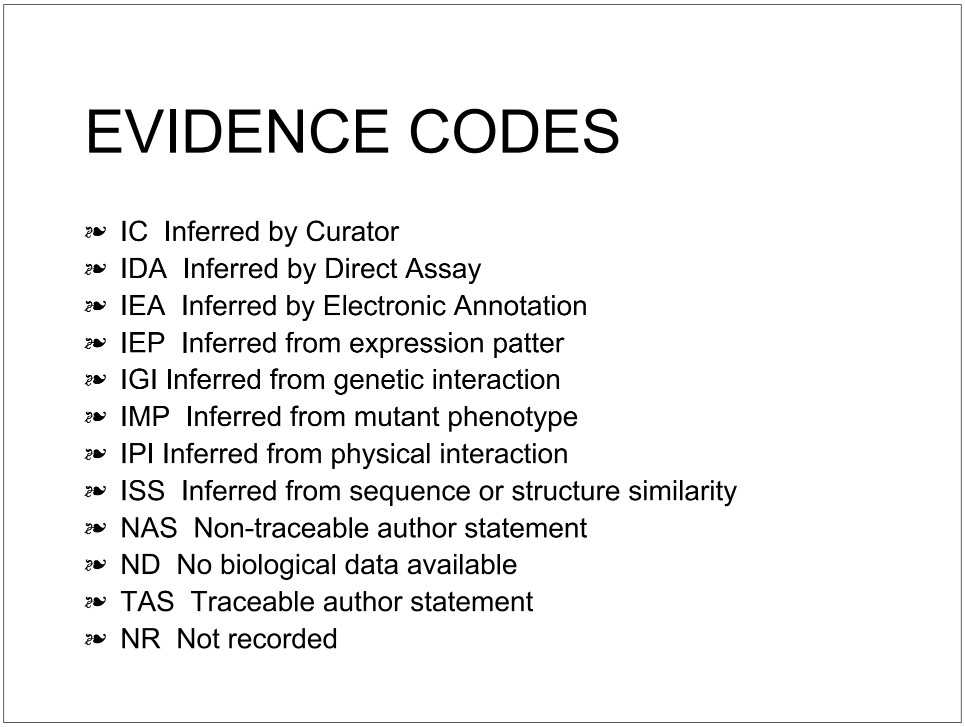
blood coagulation factor

complement activity

serine-type peptidase

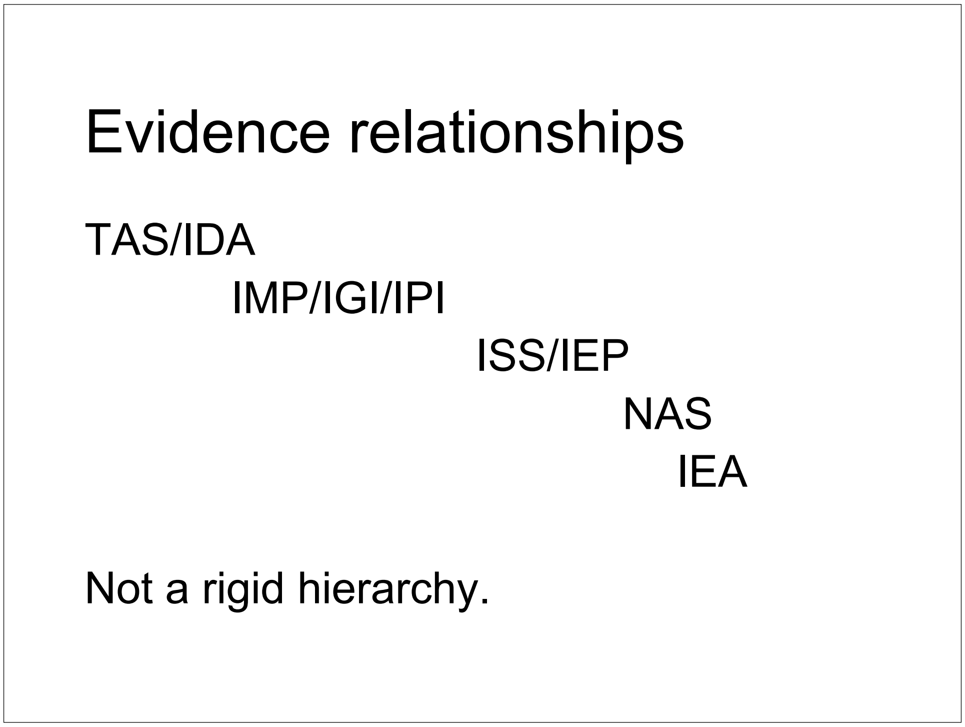
serine-type endopeptidase

blood coagulation factor VII



EVIDENCE CODES

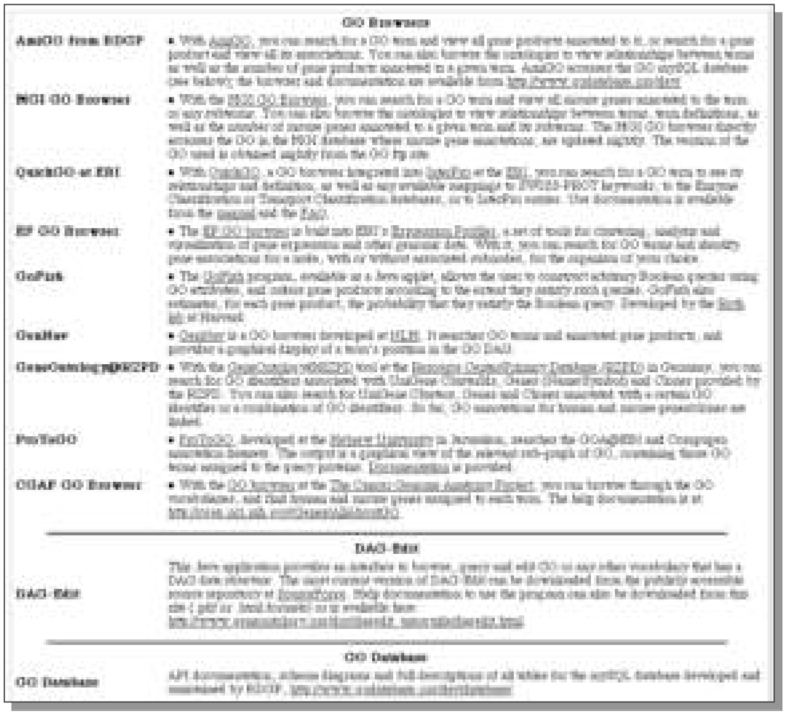
- IC Inferred by Curator
- IDA Inferred by Direct Assay
- IEA Inferred by Electronic Annotation
- IEP Inferred from expression patter
- IGI Inferred from genetic interaction
- IMP Inferred from mutant phenotype
- IPI Inferred from physical interaction
- ISS Inferred from sequence or structure similarity
- NAS Non-traceable author statement
- ND No biological data available
- TAS Traceable author statement
- NR Not recorded



Evidence relationships

TAS/IDA
IMP/IGI/IPI
ISS/IEP
NAS
IEA

Not a rigid hierarchy.



GO Browsers

AmiGO from BDGP
- With AmiGO, you can search for a GO term and view all gene products annotated to it, or search for a gene product and view all its associations. You can also browse the ontologies to view relationships between terms as well as the number of gene products annotated to a given term. AmiGO accesses the GO mySQL database (see below); the browser and documentation are available from http://www.godatabase.org/dev

MGI GO Browser
- With the MGI GO Browser, you can search for a GO term and view all mouse genes annotated to the term or any subterm. You can also browse the ontologies to view relationships between terms, term definitions, as well as the number of mouse genes annotated to a given term and its subterms. The MGI GO browser directly accesses the GO in the MGI database where mouse gene annotations are updated nightly. The version of the GO used is obtained nightly from the GO ftp site.

QuickGO at EBI
- With QuickGO, a GO browser integrated into InterPro at the EBI, you can search for a GO term to see its relationships and definition, as well as any available mappings to SWISS-PROT keywords, to the Enzyme Classification or Transport Classification databases, or to InterPro entries. Use documentation is available from the manual and the FAQ.

EP GO Browser
- The EP GO browser is built into EBI's Expression Profiler, a set of tools for clustering, analysis and visualization of gene expression and other genomic data. With it, you can search for GO terms and identify gene associations for a node, with or without associated subnodes, for the organism of your choice.

GoFish
- The GoFish program, available as a Java applet, allows the user to construct arbitrary Boolean queries using GO attributes, and orders gene products according to the extent they satisfy such queries. GoFish also estimates, for each gene product, the probability that they satisfy the Boolean query. Developed by the Roth lab at Harvard.

GenNav
- GenNav is a GO browser developed at NLM. It searches GO terms and annotated gene products, and provides a graphical display of a term's position in the GO DAG.

GeneOntology@RZPD
- With the GeneOntology@RZPD tool at the Resource Center/Primary Database (RZPD) in Germany, you can search for GO identifiers associated with UniGene Clusters, Genes (Name/Symbol) and Clones provided by the RZPD. You can also search for UniGene Clusters, Genes and Clones associated with a certain GO identifier or a combination of GO identifiers. So far, GO annotations for human and mouse genes/clones are linked.

ProToGO
- ProToGO, developed at the Hebrew University in Jerusalem, searches the GO/SwissProt and Compugen annotation datasets. The output is a graphical view of the relevant sub-graph of GO, containing those GO terms assigned to the query proteins. Documentation is provided.

CGAP GO Browser
- With the GO browser at the The Cancer Genome Anatomy Project, you can browse through the GO vocabularies, and find human and mouse genes assigned to each term. The help documentation is at http://cgap.nci.nih.gov/Genes/GOBrowserHelp.

DAG-Edit

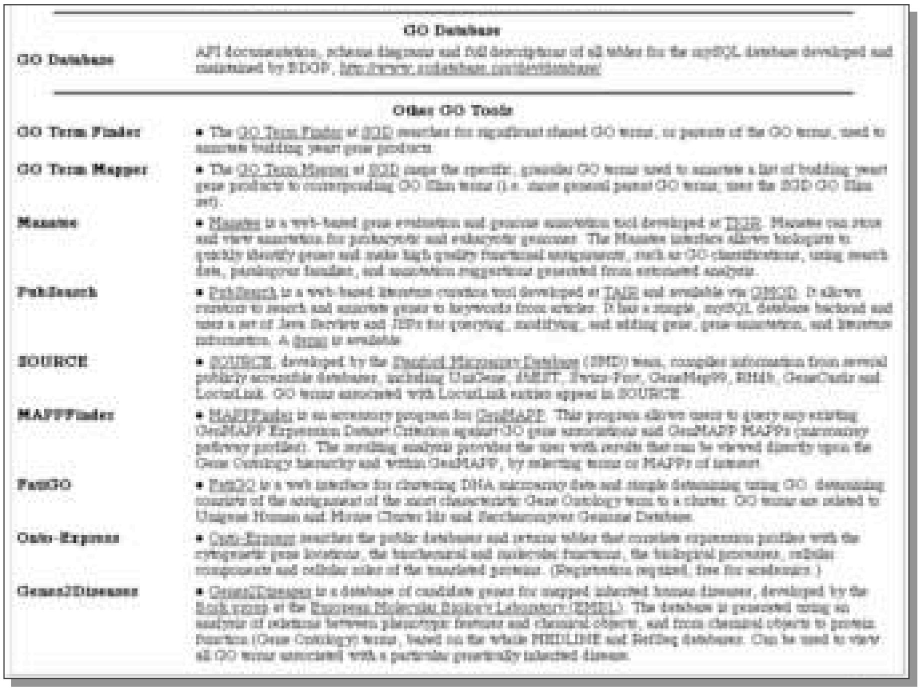
DAG-Edit

This Java application provides an interface to browse, query and edit GO or any other vocabulary that has a DAG data structure. The most current version of DAG-Edit can be downloaded from the publicly accessible source repository at SourceForge. Help documentation to use the program can also be downloaded from this site (.pdf or .html formats) or is available here: http://www.geneontology.org/doc/dagedit_user/dagedit.html

GO Database

GO Database

API documentation, schema diagrams and full descriptions of all tables for the mySQL database developed and maintained by BDGP, http://www.godatabase.org/dev/database/



GO Database

GO Database

API documentation, schema diagrams and full descriptions of all tables for the mySQL database developed and maintained by BDGP, http://www.godatabase.org/dev/database/

Other GO Tools

GO Term Finder
- The GO Term Finder at SGD searches for significant shared GO terms, or parents of the GO terms, used to annotate budding yeast gene products.

GO Term Mapper
- The GO Term Mapper at SGD maps the specific, granular GO terms used to annotate a list of budding yeast gene products to corresponding GO Slim terms (i.e. more general parent GO terms; uses the SGD GO Slim set).

Manatee
- Manatee is a web-based gene evaluation and genome annotation tool developed at TIGR. Manatee can store and view annotation for prokaryotic and eukaryotic genomes. The Manatee interface allows biologists to quickly identify genes and make high quality functional assignments, such as GO classifications, using search data, paralogous families, and annotation suggestions generated from automated analysis.

PubSearch
- PubSearch is a web-based literature curation tool developed at TAIR and available via GMOD. It allows curators to search and annotate genes to keywords from articles. It has a simple, mySQL database backend and uses a set of Java Servlets and JSPs for querying, modifying, and adding gene, gene-annotation, and literature information. A demo is available.

SOURCE
- SOURCE, developed by the Stanford Microarray Database (SMD) team, compiles information from several publicly accessible databases, including UniGene, dbEST, Swiss-Prot, GeneMap99, RHdb, GeneCards and LocusLink. GO terms associated with LocusLink entries appear in SOURCE.

MAPPFinder
- MAPPFinder is an accessory program for GenMAPP. This program allows users to query any existing GenMAPP Expression Dataset Criterion against GO gene associations and GenMAPP MAPPs (microarray pathway profiles). The resulting analysis provides the user with results that can be viewed directly upon the Gene Ontology hierarchy and within GenMAPP, by selecting terms or MAPPs of interest.

FatiGO
- FatiGO is a web interface for clustering DNA microarray data and simple datamining using GO, determining [illegible] of the assignment of the most characteristic Gene Ontology term to a cluster. GO terms are related to Unigene Human and Mouse Cluster Ids and Saccharomyces Genome Database.

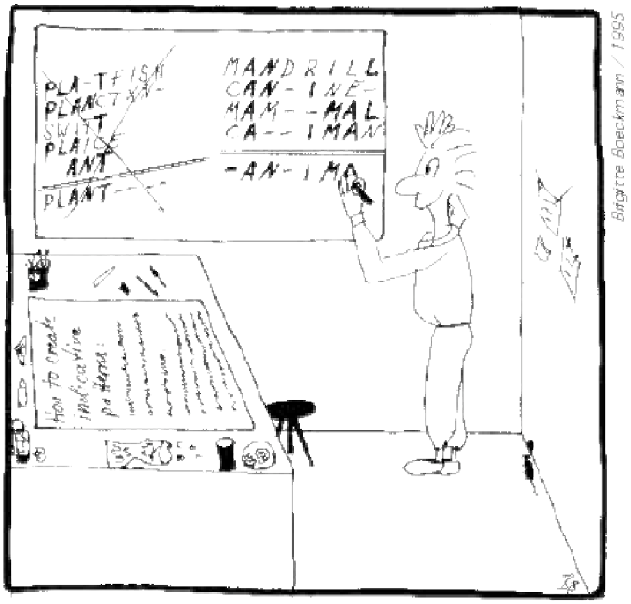
Cyto-Express
- Cyto-Express searches the public databases and returns tables that correlate expression profiles with the cytogenetic gene locations, the biochemical and molecular functions, the biological processes, cellular components and cellular roles of the translated proteins. (Registration required, free for academics.)

Genes2Diseases
- Genes2Diseases is a database of candidate genes for mapped inherited human diseases, developed by the Bork group at the European Molecular Biology Laboratory (EMBL). The database is generated using an analysis of relations between phenotypic features and chemical objects, and from chemical objects to protein function (Gene Ontology) terms, based on the whole MEDLINE and RefSeq databases. Can be used to view all GO terms associated with a particular genetically inherited disease.

Other Resources

- NCBI Education Page
  http://www.ncbi.nlm.nih.gov/Education/index.html
- BCM Gene Finder
  http://searchlauncher.bcm.tmc.edu/docs/sl_links.html
- EBI-SwissProt, TrEMBL, PIR, SRS, Tools http://www.ebi.ac.uk
- ExPASy-SwissProt, TrEMBL
  http://www.expasy.ch/
- DISC-DNA Information and Stock Center http://www.dna.affrc.go.jp

How we develop Prosite patterns!



"'Potential', 'Probable', 'By Similarity' –
They don't know anything about this protein!"